Hepburn, A., Laparra, V., McConville, R., & Santos-Rodriguez, R. (2019). Enforcing Perceptual Consistency on Generative Adversarial Networks by Using the Normalised Laplacian Pyramid Distance. Manuscript submitted for publication.

Early version, also known as pre-print

Link to publication record in Explore Bristol Research

PDF-document

This is the submitted manuscript (SM). It first appeared online via Arxiv at https://arxiv.org/abs/1908.04347. Please refer to any applicable terms of use of the author.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms arXiv:1908.04347v1 [cs.CV] 9 Aug 2019

Enforcing Perceptual Consistency on Generative Adversarial Networks by Using the Normalised Laplacian Pyramid Distance

Alexander Hepburn[1], **Valero Laparra**[2], **Ryan McConville**[1], **Raul Santos-Rodriguez**[1]
[1]Engineering Mathematics, University of Bristol
[2]Image and Signal Processing Group, Universitat de València
alex.hepburn@bristol.ac.uk, valero.laparra@uv.es, ryan.mcconville@bristol.ac.uk, enrsr@bristol.ac.uk

Abstract

In recent years there has been a growing interest in image generation through deep learning. While an important part of the evaluation of the generated images usually involves visual inspection, the inclusion of human perception as a factor in the training process is often overlooked. In this paper we propose an alternative perceptual regulariser for image-to-image translation using conditional generative adversarial networks (cGANs). To do so automatically (avoiding visual inspection), we use the Normalised Laplacian Pyramid Distance (NLPD) to measure the perceptual similarity between the generated image and the original image. The NLPD is based on the principle of normalising the value of coefficients with respect to a local estimate of mean energy at different scales and has already been successfully tested in different experiments involving human perception. We compare this regulariser with the originally proposed L1 distance and note that when using NLPD the generated images contain more realistic values for both local and global contrast. We found that using NLPD as a regulariser improves image segmentation accuracy on generated images as well as improving two no-reference image quality metrics.

Introduction

Recently, deep learning methods have become state-of-the-art in conditional and unconditional image generation (Radford, Metz, and Chintala 2016; Odena, Olah, and Shlens 2017), achieving great success in numerous applications. Image-to-image translation is one such application, where the task involves the translation of one scene representation into another representation. It has been shown that neural network architectures are able to generalise to different datasets and learn various translations between scene representations. Further, semantic labels have been used to generate realistic looking scenes which can then be used for data augmentation, e.g., in an autonomous car system (Isola et al. 2017), where new scenes can be generated by handcrafted semantic label maps.

Most state of the art methods in image-to-image translation typically use a Generative Adversarial Network (GAN) loss with regularisation. The aim of this regularisation is to maintain the overall structure of the input image in the output image. This is typically achieved with functions such as the L1, L2 or mean squared error (MSE). However, these do not account for the human visual system's perception of quality. For example, the L1 loss uses a pixel to pixel similarity which fails to capture the global or local structure of the image.

The main objective of these methods is to generate images that look *perceptually* indistinguishable from the training data to humans. Despite this, metrics which attempt to capture different aspects of images that are important to humans are ignored. Although neural networks seem to transform the data to a domain where the Euclidean distance induce a spatially invariant image similarity metric, given a diverse enough training dataset (Zhang et al. 2018), we believe that explicitly including key attributes of human perception is an important step when designing similarity metrics for image generation.

Therefore, in this paper we propose the use of a perceptual distance measure based on the human visual system that encapsulates the structure of the image at various scales, whilst normalising locally the energy of the image; the Normalised Laplacian Pyramid Distance (NLPD). This distance was found to correlate with human perceptual quality when images are subjected to perturbations such as Gaussian noise, mean shift and compression (Laparra et al. 2016). NLPD has been shown to be superior in predicting human perceptual similarity, compared to a number of well-known metrics such as the MS-SSIM (Wang, Simoncelli, and Bovik 2003) and MSE.

The main contributions of this paper are as follows:

- We argue that human perception should be used in the objective function of cGANs.
- We propose a regulariser for cGANs that measures human perceptual quality in the form of NLPD.
- We evaluate our proposed method, comparing it with the L1 loss using no-reference image quality metrics, image segmentation accuracy and an Amazon Mechanical Turk survey.
- We show improved performance over L1 regularisation, demonstrating the benefits of an image quality metric inspired by the human visual system in the objective function.

Related Work

Previously, image-to-image translation systems have been designed by experts and can only be applied to their respective representations, while being unable to learn different translations (Hertzmann et al. 2001; Chen et al. 2009). Neural network are often able to generalise and learn a variety of mappings and have proven to be successful in image generation (Radford, Metz, and Chintala 2016).

Conditional Generative Adversarial Networks

Generative Adversarial Networks (GANs) aim to generate data indistinguishable from the training data (Goodfellow et al. 2014). The generator network G learns a mapping from noise vector z to target data y, $G(z) \rightarrow y$ and the discriminator network D learns mapping from data x to label $[0, 1]$, $D(x) \rightarrow [0, 1]$ corresponding to whether the data is real or generated. GANs have become very successful in complex tasks such as image generation (Radford, Metz, and Chintala 2016). Conditional GANs (cGANs) aim to learn a generative model that will sample data according to some attribute e.g. 'generate data from class A' (Mirza and Osindero 2014). This attribute is used to build a conditional generative model where the generator generates the data with respect to the attribute and the discriminator predicts whether the data is real or generated subject to the attribute.

LAPGAN

Laplacian Pyramid Generative Adversarial Networks (LAPGANs) (Denton et al. 2015) use the laplacian pyramid network framework in order to generate images of increasing resolution. At each stage of the pyramid, a separate GAN is trained to generate a higher resolution image, given the output of the previous stage. Although this algorithm uses the underlying framework, the method is vastly different to what is proposed in this paper. Training a GAN at each stage of a laplacian pyramid requires a large amount of parameters and computation time and given that GANs are troublesome to train on their own, training a cascade of GANs is extremely time consuming. As such we suggest the use of a similar loss function, using only a single GAN and with an additional normalisation step at each stage of the pyramid. This reduces the number of parameters and computation time massively.

pix2pix

One application of cGANs is image-to-image translation, where the generator is conditioned on an input image to generate a corresponding output image (Isola et al. 2017). Isola *et al.* proposed that the cGAN objective function has a structured loss, whereby the GAN considers the structure of the output space and pixels are conditionally-dependent on all other pixels in the image.

Optimising for the GAN objective alone creates images that lack outlines for the objects in the semantic label map and a common practice is to use either the L2 or L1 loss as a reconstruction loss. Isola *et al.* preferred the L1 loss, finding that the L2 loss encouraged smoothing in the generated images. The L1 loss is a pixel level similarity metric, meaning it only cares about the distance between single pixel values ignoring the local structure that could capture perceptual similarity.

Further using a related method, it has been shown that the style of one image can be changed to match the style of a specified image (Zhu et al. 2017). CycleGAN is an extension of pix2pix where image-to-image translation is performed bidirectionally and the distance between ground truth images and images that have been translated to the other domain then translated back is calculated and used in the objective function. As a form of regularisation, a loss is introduced that aims to measure perceptual similarity often called the Visual Geometry Group (VGG) network loss.

Perceptual Distances

When the output of a machine learning algorithm will be evaluated by human observers, the image quality metric (IQM) used in the optimisation objective should take into account human perception.

In the deep learning community, the VGG loss (Dosovitskiy and Brox 2016) has been used to address the issue of generating images using perceptual similarity metrics. This method relies on using a network trained to predict perceptual similarity between two images. It has been shown to be robust to small structural perturbations, such as rotations, which is a downfall of more traditional image quality metrics such as the structural similarity index (SSIM). However, the architecture design and the optimisation takes no inspiration from human perceptual systems and treats the problem as a simple regression task; given image A and image B, output a similarity that mimics the human perceptual score.

There is a long tradition of IQMs based on human perception. Probably the most well know is the SSIM or its multi scale version (MS-SSIM) (Wang, Simoncelli, and Bovik 2003). While these distances focus on predicting the human perceptual similarity, their formulation is disconnected from the processing pipeline followed by the human visual system. On the contrary, metrics like the one proposed in by Laparra *et al.* are inspired by the early stages of the human visual cortex and show better performance in mimicking human perception than SSIM and MS-SSIM in different human rated databases (Laparra, Muñoz-Marí, and Malo 2010). In this work we use an improved version of this metric, the Normalised Laplacian Pyramid Distance (NLPD), proposed by Laparra *et al.* (Laparra et al. 2016).

Normalised Laplacian Pyramid

The Laplacian Pyramid is a well known image processing algorithm for image compression and encoding (Burt and Adelson 1983). The image is encoded by performing convolutions with a low-pass filter and then subtracting this from the original image multiple times, each time downsampling the image. The resulting filtered versions of the image have low variance and entropy and as such can be expressed with less storing information.

Normalised Laplacian Pyramid (NLP) extends the Laplacian pyramid with a local normalisation step on the output of each stage. These two steps are similar to the early stages of the human visual system. Laparra *et al.* proposed an IQM

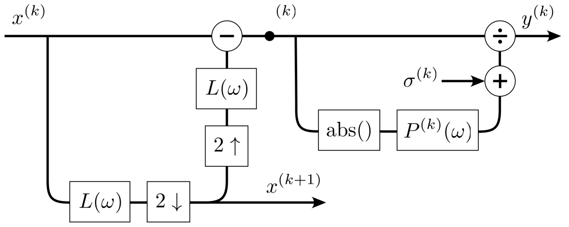


Figure 1: Figure taken from (Laparra et al. 2016). Architecture for one stage k of the Normalised Laplacian Pyramid model, where $x^{(k)}$ is the input at stage k, $L(\omega)$ is a convolution with a low-pass filter, $[2\downarrow]$ is a downsample by factor two, $[2\uparrow]$ is an upsample of factor two, $x^{(k+1)}$ is the input image at stage $(k+1)$, $P^{(k)}(\omega)$ is s scale-specific filter for normalising the image with respect to the local amplitude, $\sigma^{(k)}$ is scale-specific constant and $y^{(k)}$ is the output at scale k. The input image is defined as $x^{(0)}$.

based on computing distances in the NLP transformed domain, the NLPD (Laparra et al. 2016). It has been shown that NLPD correlates better with human perception than the previously proposed IQMs. NLPD has been employed successfully to optimise image processing algorithms, for instance to design an image compression algorithm (Ballé, Laparra, and Simoncelli 2016) and to perceptually optimised image rendering processes (Laparra et al. 2017). It has also been shown that the NLP reduces the correlation and mutual information between the image coefficients, which is in agreement with the efficient coding hypothesis (Barlow 1961), proposed as a principle followed by the human brain.

Specifically NLPD uses a series of low-pass filters, downsampling and local energy normalisation to transform the image into a 'perceptual space'. A distance is then computed between two images within this space. The normalisation step divides by a local estimate of the amplitude. The local amplitude is a weighted sum of neighbouring pixels where the weights are pre-computed by optimising a prediction of the local amplitude using undistorted images from a different dataset. The downsampling and normalisation are done at N stages, a parameter set by the user. An overview of the architecture is detailed in Figure (1).

After computing each $y^{(k)}$ output at every stage of the pyramid, the final distance is the root mean square error between the outputs of two images:

$$\mathcal{L}_{NLPD} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{\sqrt{N_s^{(k)}}}||y_1^{(k)} - y_2^{(k)}||_2, \qquad (1)$$

where N is the number of stages in the pyramid, $N_s^{(k)}$ is the number of coefficients at stage k, $y_1^{(k)}$ is the output at stage k when the input is a training image and $y_2^{(k)}$ is the output at stage k when the input is a generated image.

Qualitatively, the transformation to the perceptual space defined by NLPD transforms images such that the local contrast is normalised by the contrast of each pixels neighbours. This leads to NLPD heavily penalising differences in local contrast. Using NLPD as a regulariser enforces a more realistic local contrast and, due to NLPD observing multiple resolutions of the image, it also improves global contrast

In image generation, perceptual similarity is the overall goal; fooling a human into thinking a generated image is real. As such, NLPD would be an ideal candidate regulariser for generative models, GANs in particular.

NLPD as a Regulariser

For cGANs, the objective function is given by

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D) =& \mathbb{E}_{x,y}[\log D(x,y)] + \\ & \mathbb{E}_{x,z}[\log(1 - D(G(x,z))]\end{aligned} \qquad (2)$$

where G maps image x and noise z to target image y, $G: x, z \to y$ and D maps image x and target image y to a label in $[0, 1]$.

With the L1 regulariser proposed by Isola *et al.* (Isola et al. 2017) for image-to-image translation, this becomes

$$\mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}, \qquad (3)$$

where $\mathcal{L}_{L1} = \mathbb{E}_{x,y,z}[||y - G(x,z)||_1]$ and λ is a tunable hyperparameter.

In this paper we propose replacing the L1 regulariser $\mathcal{L}_{L1}$ with a NLPD regulariser. In doing so the entire objective function is given by

$$\mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{NLPD}. \qquad (4)$$

In the remainder of the paper Eq. (3) will be denoted by cGAN+L1 and Eq. (4) by cGAN+NLPD.

Computation Time

NLPD involves 3 convolution operations per stage in the pyramid, with the same convolution applied independently to each colour channel of the input. Although this is more computationally expensive than $L1$ loss, relative to the entire training procedure of training a GAN, the increase in computation time is negligible.

In addition to this, with computational packages like Tensorflow and Pytorch, the process of transforming images into the perceptual space via a laplacian pyramid can simply be appended to the generator computation graph as extra convolutional layers with a very low number of parameters compared to traditional convolutional layers. There are $3 \times k$ convolution filters, where k is the number of stages in the pyramid, that should be stored in memory but the number of filters stored in a network is several orders of magnitude greater.

Experiments

Datasets

We evaluated our method on three public datasets, each varying in difficulty and subject matter; the Facades dataset (Tyleček and Šára 2013), the Cityscapes dataset (Cordts et al. 2016) and a Maps dataset (Isola et al. 2017). Colour images were generated from semantic label maps for both the Facades dataset and the Cityscapes dataset. The Facades dataset is a set of architectural label drawings and the corresponding colour image for various buildings. The Cityscapes dataset is a collection of label maps and colour images taken from the a front facing car camera, as it drives around various cities. For the Cityscapes dataset, images were resized to a resolution of 256×256 and after generating the images they were resized to the original dataset aspect ratio of 512×256, as the network architecture used works best on square images. The third dataset is a Maps dataset of images taken from Google Maps that was constructed by Isola *et al.*. It contains a map layout image of an area and the corresponding aerial image resized to a resolution of 256×256.

The objective of all of these tasks is to generate a RGB image from the textureless label map. For all datasets, the same train and test splits were used as in the pix2pix paper, in order to ensure a fair comparison.

Experimental Setup

For all experiments, the architecture of both the generator and discriminator is the same as defined by Isola *et al.* (Isola et al. 2017). The generator is a U-net with skip connections between each mirroring layer. The discriminator is a patch discriminator which observes 70×70 pixel patches at a time, with dropout applied at training. Full architecture can be found in the paper by Isola *et al.* or in the pix2pix repository [1]. In our method we use the least-squares adaptation of the GAN loss as it improves stability (Mao et al. 2017). We also used the Adam optimiser (Kingma and Ba 2014) with learning rate 0.0002 and trained each network for 200 epochs. A batch-size of 1 was used with batch normalisation and each layer had ReLU activations applied to them. This methodology is essentially using an instance normalisation layer (Ulyanov, Vedaldi, and Lempitsky 2017) and has been found to be ideal in training image-to-image translation models (Isola et al. 2017). Random cropping and mirroring were applied during training.

For the L1 regulariser, a λ value of 100 was used, the optimal value found by Isola *et al.* (Isola et al. 2017). For NLPD, $\lambda = 15$ was found to be best after a hyperparameter search. The number of stages was chosen as $N = 6$ ensuring that at the final stage the resolution of the output image will be 4×4. The normalisation filters were found by optimising the weights to recover the original local amplitude from various perturbed images using the McGill dataset (Olmos and Kingdom 2004). As these weights were found by optimising over black and white images, we apply the normalisation to each channel independently.

We vary the objective function that the network is trained with in order to highlight the effect of including the Normalised Laplacian Pyramid Distance as a regulariser.

Evaluation

Evaluating generative models is a difficult task (Theis and Bethge 2015). Therefore we have performed different experiments to illustrate the improvement in the performance when using NLPD as regulariser. In image-to-image translation, there is additional information in the form of the label map that images were generated with. A common metric involves evaluating how well a network trained on the ground truth performs at a task such as image segmentation on the generated images (Isola et al. 2017; Wang et al. 2018). Naturally, generated images which achieve higher performance at this task can be considered more realistic. One architecture that has been successfully used for image segmentation is the fully convolution network (FCN) (Long, Shelhamer, and Darrell 2015).

FCN-Score In traditional image classification networks, the final layers often involve fully connected layers. FCNs replace these fully connected layers with fully convolutional layers to represent label heat maps (Long, Shelhamer, and Darrell 2015). As such, most image classification networks can be adapted into image segmentation networks.

We use the typical approach from the literature (Isola et al. 2017) and train a FCN-8 for image segmentation on the Cityscapes dataset at a 256×256 resolution. Generated images are then produced from label maps in the validation set of the Cityscapes dataset. Following this, 3 image segmentation accuracy metrics are calculated. Per-pixel accuracy is the percentage of pixels correctly classified, per-class accuracy is the mean of the accuracies for all classes and class IOU is the intersection over union, which measures the percentage overlap between the ground truth label map and the predicted one.

We note that the ground truth accuracy is lower due to the network being trained on images of resolution 256×256, which are then upsampled to the full resolution of the label map, 2048×1024.

Loss	**Pre-Pixel Accuracy**	**Per-Class Accuracy**	**Class IOU**
cGAN+L1	0.71	0.25	0.18
cGAN+NLPD	0.74	0.25	0.19
Ground Truth	0.80	0.26	0.21

Table 1: FCN-scores for each loss function trained on Cityscapes label→photo. In cGAN+NLPD $\lambda = 15$, and in cGAN+L1 $\lambda = 100$.

No-Reference Image Quality Metrics Traditional image quality metrics often require a reference image, e.g., measuring the root mean square error between a generated image and the ground truth. However, when generating an image from a label map, the ground truth is just one possible solution.

There exist many images that could be feasibly generated from one label map and, as such, reference image quality metrics are unsuitable. Therefore we include two no-reference image quality metrics to more thoroughly evaluate the generated images, namely BRISQUE and NIQE.

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) is an image quality metric that aims to mea-

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

sure the 'naturalness' of an image using statistics of locally normalised luminance coefficients (Mittal, Moorthy, and Bovik 2012). For natural images, these coefficients normally follow a Gaussian distribution (Ruderman 1994) and BRISQUE measures how well the mean subtracted contrast normalised (MSCN) coefficients fit a generalised Gaussian distribution. BRISQUE also measures how well a set of pairwise products between four orientations of the MSCN image fit an asymmetric generalised Gaussian distribution. The four orientations are vertical, horizontal, right-diagonal and left diagonal in order to capture the relationship between a pixel and it's neighbours. Overall, BRISQUE was found to be an improvement over some full-reference image quality metrics, e.g., the structural scale similarity (SSIM).

Natural Image Quality Evaluator (NIQE) (Mittal, Soundararajan, and Bovik 2013) is a fully blind image quality metric in that it has no knowledge of the types of distortions applied to the images. NIQE selects patches of the image that provide the most information and computes statistics such as local variance inside the set of patches. The distribution of these statistics for a query image is then compared to the distribution of natural images and a score is calculated.

Amazon Mechanical Turk As our objective is to generate images which, to humans, look perceptually similar to the original images, we also evaluate the performance by asking humans to judge the quality of the generated images.

Experiments were conducted using Amazon Mechanical Turk (AMT) and users were asked to chose "Which image looks more natural?" when presented with one image generated using the L1 regulariser and another by NLPD regulariser. A random subset of 100 images were chosen from the validation set of each dataset and 5 unique decisions were gathered per image. The placement on the left or right of the images for each regulariser were randomly permuted.

Results

Results of images generated using the proposed procedure and the L1 baseline for the three different datasets are presented in Figs. 3a, 3b, and 2.

Loss Function	BRISQUE(NIQE) Scores		
	Facades	**Cityscapes**	**Maps**
cGAN+L1	30.08 (5.23)	26.57 (3.86)	30.63 (4.71)
cGAN+NLPD	30.06 (5.21)	24.54 (3.57)	28.99 (4.59)
Ground Truth	37.29 (7.33)	25.40 (3.12)	28.48 (3.35)

Table 2: BRISQUE and NIQE scores for various datasets and loss functions. For both, the lower the score, the more natural the image is.

Table 1 shows results for the FCN-scores for the images generated using the Cityscapes database. In general the images generated using NLPD show improvement over the L1 regularisation, in particular in the per-pixel accuracy and class IOU. As such, it can be seen that the NLPD images contain more features of the original dataset according to the FCN image segmentation network.

Table 2 shows the scores for both the BRISQUE and NIQE image quality metrics. The two no-reference image quality metrics aim to measure the naturalness of an image. A lower value means a more natural image. On average, NLPD regularisation achieves lower values in both metrics. For Cityscapes and Maps, NLPD is close to the scores achieved by the ground truth. The ground truth scores for the Facades dataset can be worse than the generated images due to the large grey or black triangles that are in the Facades training set, included to crop out some of the sky and neighbouring buildings. These triangles are very unnatural textures and as such could cause the scores to be significantly worse.

Using Amazon Mechanical Turk we tested the human perceived quality by querying users regarding the naturalness of the presented images. The percentage of users that found the NLPD images more natural was above chance for the Maps (52.37%) and Cityscapes datasets (56.16%), while similar for Facades (50.04%). Visual inspection of Fig. 3a shows that when generating from a map that contains a large building, NLPD produces more realistic textures, whereas L1 contains repeating patterns. In the Cityscapes dataset the contrast appears slightly more realistic, e.g., the white in the sky is lighter in Fig. 2, which could result in users preferring these images. In images generated using the Facades dataset, it is hard to visually find differences in Fig. 3b and therefore difficult to measure a preference between the two regularisers.

Conclusion

Taking into account human perception in machine learning algorithms is challenging and usually ignored in automatic image generation. In this paper we detailed a procedure to take into account human perception in a conditional GAN framework. We propose to modify the standard objective by incorporating a term that accounts for perceptual quality by using the Normalised Laplacian Pyramid Distance (NLPD). We illustrate its behaviour in the image-to-image translation task for a variety of datasets. The suggested objective shows better performance in all the evaluation procedures. Interestingly, it also has a better segmentation accuracy using a network trained on the original dataset, and produces more natural images according to two no-reference image quality metrics. In human perceptual experiments, users showed a preference for the images generated using the NLPD regulariser over those generated using L1 regularisation.

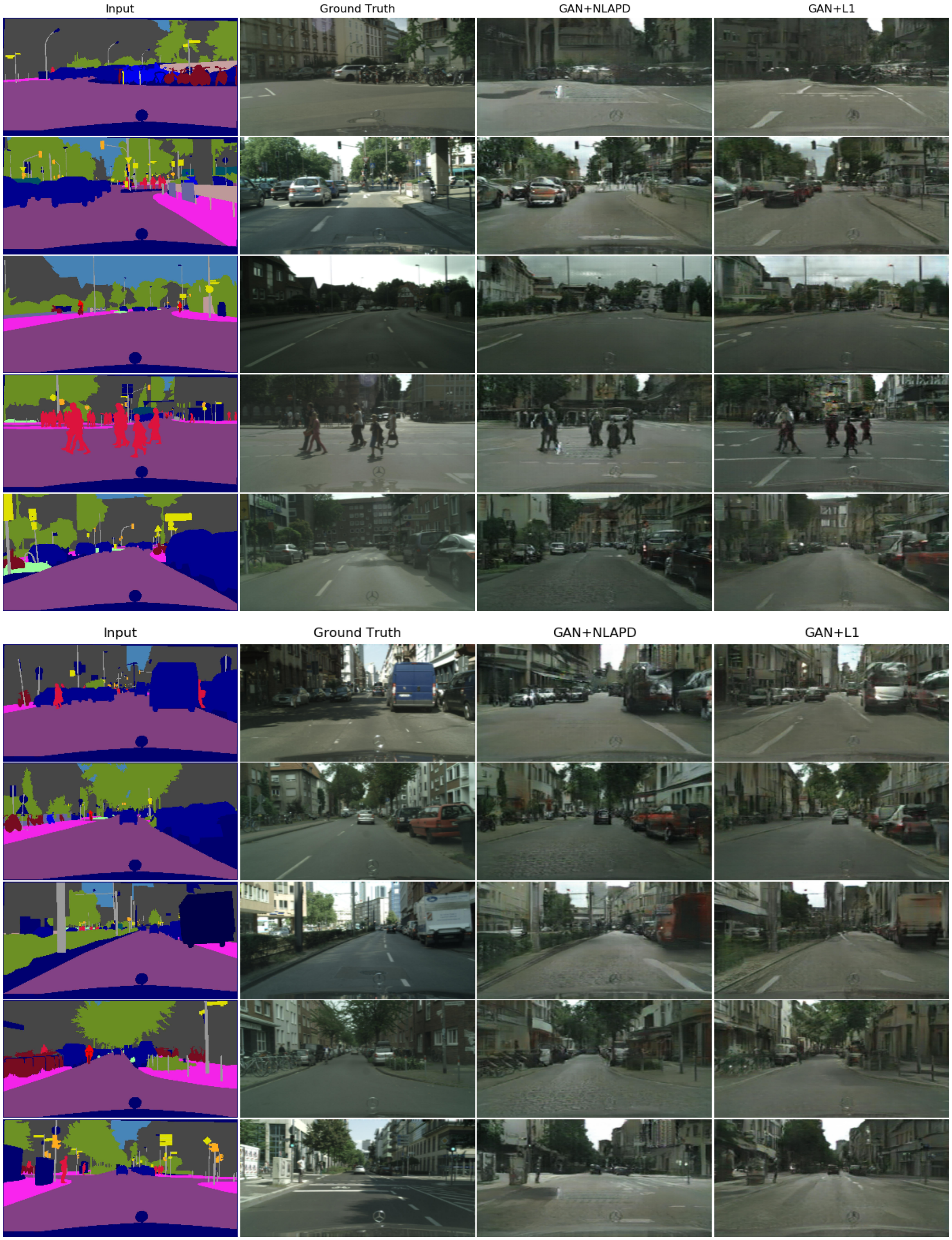
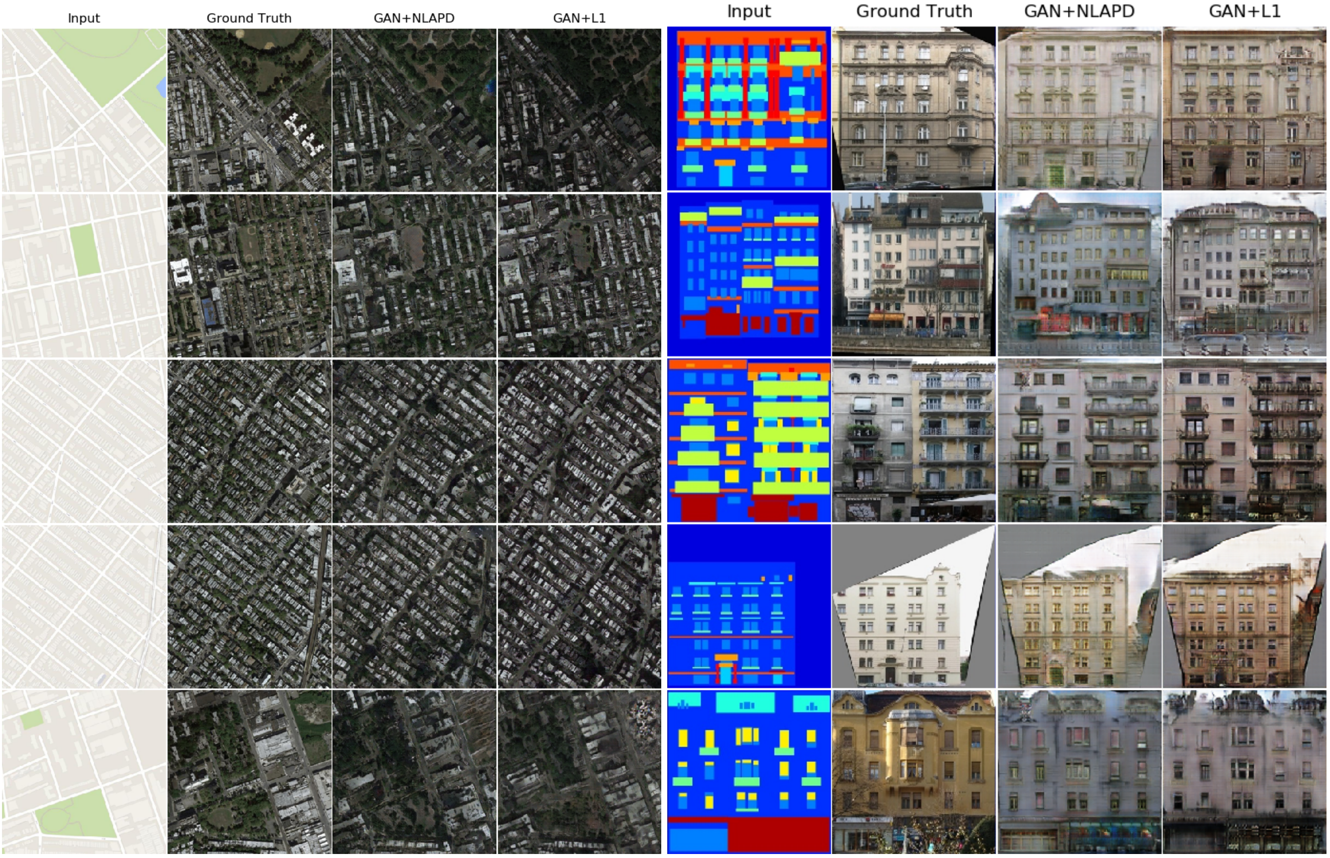


Figure 2: Images generated from label maps taken from the Cityscapes validation set. Images were generated at a resolution of 256×256 and then resized to the original aspect ratio of 512×256.



(a) Maps

(b) Facades

Figure 3: Images generated from the (a) Maps dataset and (b) Facades dataset at a resolution of 256×256 using both L1 and NLPD regularisation.

References

[Ballé, Laparra, and Simoncelli 2016] Ballé, J.; Laparra, V.; and Simoncelli, E. P. 2016. End-to-end optimization of nonlinear transform codes for perceptual quality. In *Proceedings of the PCS*.

[Barlow 1961] Barlow, H. B. 1961. Possible principles underlying the transformation of sensory messages. *Sensory Communication* 217–234.

[Burt and Adelson 1983] Burt, P., and Adelson, E. 1983. The laplacian pyramid as a compact image code. *IEEE Transactions on communications* 31(4):532–540.

[Chen et al. 2009] Chen, T.; Cheng, M.; Tan, P.; Shamir, A.; and Hu, S. 2009. Sketch2photo: Internet image montage. In *ACM transactions on graphics*, volume 28, 124.

[Cordts et al. 2016] Cordts, M.; Omran, M.and Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE CVPR*, 3213–3223.

[Denton et al. 2015] Denton, E. L.; Chintala, S.; Fergus, R.; et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, 1486–1494.

[Dosovitskiy and Brox 2016] Dosovitskiy, A., and Brox, T. 2016. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, 658–666.

[Goodfellow et al. 2014] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

[Hertzmann et al. 2001] Hertzmann, A.; Jacobs, C. E.; Oliver, N.; Curless, B.; and Salesin, D. H. 2001. Image analogies. In *Computer graphics and interactive techniques*, 327–340. ACM.

[Isola et al. 2017] Isola, P.; Zhu, J.; Zhou, T.; and Efros, A. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE CVPR*.

[Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Laparra et al. 2016] Laparra, V.; Ballé, J.; Berardino, A.; and P, S. E. 2016. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging* 2016(16):1–6.

[Laparra et al. 2017] Laparra, V.; Berardino, A.; Ballé, J.; and Simoncelli, E. P. 2017. Perceptually optimized image rendering. *Journal Optical Society of America, A*.

[Laparra, Muñoz-Marí, and Malo 2010] Laparra, V.; Muñoz-Marí, J.; and Malo, J. 2010. Divisive normalization image quality metric revisited. *Journal of the Optical Society of America A* 27(4):852–864.

[Long, Shelhamer, and Darrell 2015] Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, 3431–3440.

[Mao et al. 2017] Mao, .; Li, Q.; Xie, H.; Lau, R. Y. K.; Wang, Z.; and Smolley, S. P. 2017. Least squares generative adversarial networks. In *IEEE ICCV*, 2794–2802.

[Mirza and Osindero 2014] Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *CoRR* abs/1411.1784.

[Mittal, Moorthy, and Bovik 2012] Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* 21(12):4695–4708.

[Mittal, Soundararajan, and Bovik 2013] Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a" completely blind" image quality analyzer. *IEEE Signal Process. Lett.* 20(3):209–212.

[Odena, Olah, and Shlens 2017] Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, 2642–2651. JMLR. org.

[Olmos and Kingdom 2004] Olmos, A., and Kingdom, F. A. A. 2004. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception* 33(12):1463–1473.

[Radford, Metz, and Chintala 2016] Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR* abs/1511.06434.

[Ruderman 1994] Ruderman, D. L. 1994. The statistics of natural images. *Network: computation in neural systems* 5(4):517–548.

[Theis and Bethge 2015] Theis, Land Oord, A., and Bethge, M. 2015. A note on the evaluation of generative models. *International Conference on Learning Representations*.

[Tyleček and Šára 2013] Tyleček, R., and Šára, R. 2013. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, 364–374. Springer.

[Ulyanov, Vedaldi, and Lempitsky 2017] Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6924–6932.

[Wang et al. 2018] Wang, T.; Liu, M.; Zhu, J.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE CVPR*, 8798–8807.

[Wang, Simoncelli, and Bovik 2003] Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, volume 2, 1398–1402. Ieee.

[Zhang et al. 2018] Zhang, R.; Isola, P.; Efros, A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE CVPR*, 586–595.

[Zhu et al. 2017] Zhu, J.; Park, T.; Isola, P.; and Efros, A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE ICCV*, 2223–2232.